

Research Article

Validity of Mental Workload Measures in a Driving Simulation Environment

Francesco Galante ¹, **Fabrizio Bracco**,² **Carlo Chiorri**,² **Luigi Pariota** ¹,
Luigi Biggero,¹ and **Gennaro N. Bifulco** ¹

¹Department of Civil, Architectural and Environmental Engineering, University of Naples Federico II, 80125 Naples, Italy

²Department of Educational Sciences, University of Genoa, 16128 Genoa, Italy

Correspondence should be addressed to Luigi Pariota; luigi.pariota@unina.it

Received 30 March 2018; Revised 18 July 2018; Accepted 5 August 2018; Published 19 August 2018

Academic Editor: Ludovic Leclercq

Copyright © 2018 Francesco Galante et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Automated in-vehicle systems and related human-machine interfaces can contribute to alleviating the workload of drivers. However, each new functionality can also introduce a new source of workload, due to the need to attend to new tasks and thus requires careful testing before being implemented in vehicles. Driving simulators have become a viable alternative to on-the-road tests, since they allow optimal experimental control and high safety. However, for each driving simulator to be a useful research tool, for each specific task an adequate correspondence must be established between the behavior in the simulator and the behavior on the road, namely, the simulator absolute and relative validity. In this study we investigated the validity of a driving-simulator-based experimental environment for research on mental workload measures by comparing behavioral and subjective measures of workload of the same large group of participants in a simulated and on-road driving task on the same route. Consistent with previous studies, mixed support was found for both types of validity, although results suggest that allowing more and/or longer familiarization sessions with the simulator may be needed to increase its validity. Simulator sickness also emerged as a critical issue for the generalizability of the results.

1. Introduction

Reconciling mobility needs with efficient and more sustainable transportation, while increasing levels of road safety, is a key objective in the transportation sector. Although many environmental factors, such as roadway type, vehicle characteristics, weather, road conditions, traffic density, and flow may jeopardize road safety, a vast majority of traffic accidents are caused by, or at least can be related to, poor human performance (e.g., mainly, distraction, and risk misperception). Distraction and poor understanding of traffic situation could be generated by inefficient attentional resource allocation and workload level has been identified as one of the main factors affecting attention and performance [1].

The term workload can refer to a number of different yet related entities. It is related to an unbalance between mental and/or physical resources and task demands [2]. Taking into

account the characteristics of driving, physical workload is less relevant than the mental one, which results from mental processes when performing driving tasks, depending on the users' capabilities and the task demands. Mental workload can impair driving performance either when it is too high or when it is too low [3]. De Waard and Brookhuis [4] argued that both overload and underload could impair attention. The former can lead to distraction, diverted attention, and insufficient capacity and time for adequate information processing. Therefore, overload would affect the capacity to focus attention on right task or pieces of information. On the other hand, underload can cause reduced alertness, focused attention could be lowered, and the reaction to events could be slow or inaccurate. Indeed, car driving is a complex activity, consisting of both psychological and physiological demands, and workload appears inevitable.

Technology solutions have been proposed to reduce such a complex activity, alleviating the driver from demanding

tasks and shifting them to automation. Driving automation at different SAE (Society of Automotive Engineers) levels [5] is seen as having the potential to reduce road fatalities, in particular by mitigating physical demand or improving awareness of the driving environment. However, research has shown that making automation available to the driver may fail to alleviate the workload, or, somehow ironically, it may even introduce a further source of workload by creating additional task activities [6]. On the other hand, effective automation could perform both longitudinal and lateral control of the car, thus leaving to the driver the monitoring task. However, this condition may result in low demand for action and an underload. Drivers' attention may be diverted elsewhere and they could have troubles at refocusing on the driving task in case of emergency [7]. Any new automation device must therefore be carefully tested to adequately weigh its benefits and shortcomings [8].

On-road research on real vehicles can be costly in terms of money and driver safety and may not afford adequate experimental control and data collection [9]. Driving simulators (DSs) allow us to address many of these issues.

DSs have long been used by car manufacturers and research institutes to test users' acceptability of on-board devices and human-vehicle interfaces, and in recent years they have been increasingly employed also in earlier conception phases, where the feasibility, effectiveness, and safety of vehicle automation solutions have to be assessed. Studies based on DSs provide a virtual experimental environment which attempts to replicate the test road conditions as realistic as possible.

Noticeably, the use of simulation allows a wide range of test conditions to be prescribed and applied consistently, but DSs have been shown to have several drawbacks. These include simulator sickness (i.e., symptoms of discomfort, drowsiness, dizziness, and nausea), lack of or incomplete replication of physical sensations, user acceptance, ecological validity [10], and generalizability of results to the real world [11]. Notwithstanding these drawbacks, DSs appear to offer a valid alternative to on-road tests for the investigation of workload related to automated in-vehicle systems, if they possess adequate "functional fidelity"; i.e., they elicit individual differences in cognitive and affective functioning like those observed in real world settings [12].

That said, the aim of this work was to assess the validity of a fixed-based driving simulator for research on mental workload comparing workload levels measured in virtual environment with measurements carried out on road in similar conditions (same roads, same tasks) on the same large group of drivers. Both behavioral and subjective measures of workload were used.

The next section summarises previous studies on these topics and has been divided into two subsections: the first about mental workload measures and the second about simulator validity compared with on-road experience in the assessment of workload. Section 3 presents the experiment carried out in terms of participants, requested tasks, and used tools. Results concerning both subjective and behavioral measures are presented in Section 4 and are discussed in Section 5.

2. Background

2.1. Mental Workload Measures. Mental workload is a multifaceted psychological construct subject to different definitions [13]. One of the main approaches to mental workload assessment refers to the multiple resources model [14]. The notion of "resource" implies a limited entity, which can be devoted to some activity, while the notion of "multiple" implies that the resources could be different in their nature. According to this model, resources could be devoted to two broadly independent processing channels, visual and verbal, even though they could share some resources at the central processing stage. This explains why two tasks requiring different kinds of resources could be carried out without a significant decrease in performance [15].

Since workload is a wide construct, there are at least three kinds of indices that could be used for workload assessment: physiological measures, behavioral measures, and subjective measures [1]. Physiological indices, such as heart rate or blood pressure, have been widely adopted as reliable indicators of workload for more than two decades (e.g., [16]), also in the road safety domain [17–19]. Behavioral measures can be either direct or indirect. The former are based on techniques of direct registration of the driver's capability to perform the driving task at an acceptable level, i.e., avoiding errors in vehicle handling. Indirect measures are based on so-called secondary or double-tasks, such as the peripheral detection task (PDT [20]). In this approach the measure of workload is derived from the measurement of the effects of a primary task (e.g., driving) on a concurrent secondary task (e.g., target discrimination). This approach assumes that the secondary task performance should decrease as a function of the mental workload required by the primary task [21]. For the sake of workload measure, it is necessary to engage the drivers with two tasks that could interact in terms of processing resources. According to the multiple resources model [14], a visual-motor task (like driving) could be easily performed together with an auditory secondary task (like following instructions from a GPS). Therefore, the combination of visual and auditory tasks is good for road safety, but partially ineffective for the measurement of mental workload of the primary task, since the secondary one will load difference attentional resources [7]. For this reason, many studies based on dual-task conditions involved two visuomotor activities like driving and discriminating targets on a display [21]. Subjective measures allow drivers to directly report their experienced workload after the task. It is therefore necessary to provide the most accurate list of empirically observable variables that could tackle this complex construct. Several self-report tools have been developed and are usually customized on the specific task to be assessed [22]. Since workload is multifaceted, there is not a single approach that could tackle the complexity of the phenomenon and it may be necessary to adopt multiple methods [1]. However, several studies reported some dissociation between the results of subjective and behavioral measures [23–25]. This difference could be explained by taking into account the concept of effort, i.e., the amount of resources invested in the task. As a result, the driver could invest a lot of effort to keep the

performance at an acceptable level: the workload assessed by means of performance quality would therefore seem stable, while the self-report rating would reveal an increase in workload. Effort depends on motivation, task difficulty, and the subjective criteria of performance. The dissociation between performance and subjective measures should be higher in driving conditions where effort is relevant for maintaining an undisturbed performance level, e.g., before and after the optimal performance phase, when the workload is low and stable, and the performance is at its best [26]. Other studies demonstrated that behavioral tasks could be sensitive to peaks in workload and subjective methods would be affected by global task demands [27, 28]. As correctly pointed out by de Waard [26], this dissociation is not really a problem if we consider the multidimensional nature of workload. A disagreement between measures could indeed provide more information about the construct [29].

2.2. Simulator Validity and Mental Workload Differences between the Simulator and On-Road Experience. Traditionally, two types of validity are evaluated for driving simulators: physical validity and behavioral validity [30]. Physical validity refers to the degree to which there is an accurate correspondence of components, layout, and dynamics between a simulator and its real world counterpart. Physical validity is generally considered as being greater in high-level, moving-base simulators with more advanced configurations than mid-level or low-level fixed-base simulators [31]. Nevertheless, a high degree of physical validity is not necessarily required for gaining useful information on how an individual will act in a given situation. For instance, Reed and Green [32] found that high vs. low fidelity in visual scenes (given comparable screen size and viewing angle) does not generally seem to have a substantial impact on driving performance variables. On the other hand, behavioral validity refers to the extent to which a driver shows similar behaviors in the simulator and in the real world. Blaauw [30] indicates that the correspondence between the behavior in the simulator and the behavior on the road is commonly seen as the more important form of validity in the evaluation of specific task performance. Behavioral validity is defined in terms of absolute and relative validity [30, 31]. Absolute validity is the extent to which the numerical values obtained under simulation and on-road are the same for specified variables (such as speed), whereas relative validity entails the extent to which the numerical values are of similar magnitude and in the same direction in the two environments. For example, when the aim of the study is to test whether participants modify their speed as compensatory behavior when under a particular form of secondary task load, differences in absolute speed between simulator and on-road performance are not of substantial interest.

A large number of studies reported the validation of simulators on the basis of driving performance measures. Godley et al. [31] used driving speed behaviors while Törnros [33] referred to both speed and lateral position. Behavioral absolute validity was not achieved in either study, although similar conclusions were reached and the relative validity was established. Conversely, Kaptein et al. [34] found absolute

validity for route choice behavior, whereas absolute validity with respect to speed was obtained by Galante et al. [35] using the VERA driving simulator at the Road Safety Laboratory in Naples. Stanton et al. [36], grounding on Brown's studies [37] of dual-task methods for assessing workload, behaviorally validated their simulator not only with respect to the performance of the drivers, but also taking into account their psychophysical perception of the driving environment and the task at hand. Validation studies taking workload into account are far fewer. Johnson et al. [38] compared cardiopulmonary responses in simulated and on-road experiences and found that the time taken to complete the course, as well as the increase from the baseline to drive in all cardiopulmonary variables, were similar between simulated and on-road environments. Nevertheless, significantly greater mean and maximum heart rate values during on-road driving were observed. Wang et al. [39] compared the performance on three manual address entry methods (keypad, touch screen, and rotational controller) in on-road and simulated environments and found that measures of glance frequency, total glance duration, percent time eyes forward, initial response time, and mean task time mapped almost identically from simulation to field. The rank ordering of the effects of the three input methods was consistent across environments.

Previous studies rarely used the same group of participants and/or the same route for both the simulated and on-road driving task, thus undermining the validity of the comparisons. Moreover, the number of participants was sometimes very low, thus allowing a limited statistical power for detecting small but substantively interesting effects. In the present study we investigated the validity of a DS-based experimental environment for research on mental workload by comparing behavioral and subjective measures of workload of the same large group of participants in a simulated and in an on-road driving task on the same route.

3. Methods

3.1. Participants. An initial pool of 150 participants took part in the study. They responded to advertisements requesting volunteers for a study on driving behavior. A quota sampling was performed to draw a group of 100 participants that matched the Italian drivers' population on gender, age, and educational level. The admissible combination of the previous features allowed splitting the sample into 14 strata (Table 1). The cardinality of each stratum was fixed considering the relative incidence in the population provided by the Italian national statistics office (<http://www.istat.it/en/>), as updated to the latest available year. To fill in missing information, we also used data from the DATIS project, carried out by the Italian national health service (<http://www.iss.it/chis/?lang=2>) to define the distribution of gender in each stratum. Each participant had normal vision or corrected to normal vision. Participants' handedness was not assessed.

3.2. Driving Tasks. Each driving experiment consisted of two driving tests, one in a DS environment, and the other on the road. The on-road experiments were carried out by

TABLE 1: Stratification of the sample used in this study.

Layer	Age (years)	Gender	Educational Level [#]	Relative incidence	Sample size
1	20-24	Male	Low	0.2*0.429*1	9
2		Female		0.2*0.571*1	11
3	25-40	Male	Low	0.3*0.483*0.5	7
4			High	0.3*0.483*0.5	7
5		Female	Low	0.3*0.517*0.5	8
6			High	0.3*0.517*0.5	8
7	41-64	Male	Low	0.3*0.491*0.5	7
8			High	0.3*0.491*0.5	7
9		Female	Low	0.3*0.509*0.5	8
10			High	0.3*0.509*0.5	8
11	≥65	Male	Low	0.2*0.674*0.5	7
12			High	0.2*0.674*0.5	7
13		Female	Low	0.2*0.326*0.5	3
14			High	0.2*0.326*0.5	3

Note: [#]low = less than or equal to high school degree; high = higher than high school degree.

means of the Instrumented Vehicle (IV) belonging to the Department of Civil and Environmental Engineering of the University of Naples Federico II. It is equipped with systems for data acquisition and video capture as well as with a touch-screen panel PC that allowed the acquisition of data to be managed and monitored and the administration of the tasks used to measure mental workload (see below) while driving (for a more detailed description of the IV, see [40]). The simulator consisted of a single seat cockpit, with all the driver controls [41]. A real-time antialiased 3D graphical scene of the virtual world was visualized on three surrounding 23" monitors at a total resolution of 5040 x 1050 pixels. The total horizontal and vertical fields of view were 100 degrees and 20 degrees, respectively. A rear-view mirror was displayed on the central monitor, and side-view mirrors were displayed on the outer monitors. The frame rate was kept constant at 60Hz. The driving experience provided by the simulator was enhanced by a surrounding sound system that simulated the various sound sources (e.g., engine, wind, and tyre). Although the simulator was fixed-base, torque feedback at the steering wheel was provided and adjustable springs provided all the pedals (clutch, brake, and accelerator) with realistic force feedback. The software SCANeR™ Studio from Oktal company was used to model the driving simulator course and to manage the whole simulation phases.

The driving scenario was the same both in the road and in the virtual environment: a high-realistic replication of the tested road and its surroundings was specifically created for the DS driving tests. It consisted of a single loop over three roads near Naples:

- (1) National Highway A1 (14 Km), characterized by a dual-carriageway layout with three lanes plus a shoulder in each direction, and a design speed ranging between 80 and 120 km/h (posted speed limit 100 km/h)
- (2) National Highway A30 (30 Km), with the same characteristics as the National Highway A1

- (3) Rural Highway SS 268 "del Vesuvio" (16 Km), characterized by a single carriageway with one lane plus a shoulder in each direction at-grade intersections and a design speed between 60 and 100 km/h.

In the experiment on the road the three sections were preceded by a 10-km acclimatization section, followed by an 8-km urban path aimed at closing the loop, for a total of 78 Km. In the DS experiments the acclimatization section lasted 10 minutes and was used to familiarize participants with the simulator; of course it was not necessary to close the loop in the DS environment; then the urban path was not included in the DS scenario. Totally each driving test lasted about one hour in both the environments.

Three different driving conditions, one for each of three main sections of the loop, were programmed to occur during the drive:

- (1) In the NH A30 section (Section 1), the driver was immersed in a traffic stream that moved at about 100 km/h; this allowed us to collect natural car-following data, without engaging the driver in specific tasks.
- (2) In the NH A1 section (Section 2), the driver interacted with a confederate lead vehicle that carried out several standard manoeuvres; specifically, the driver was asked to perform three approaching manoeuvres with the leader at a constant speed of 80, 100, and 120 km/h.
- (3) In the (slower) two-lane rural highway section (Section 3), the corporate vehicle was not present and natural car-following data were collected.

The confederate vehicle used its cruise control to keep it at a constant speed on the right lane of the NH A1. It was driven by a member of the research staff in coordination with another research staff member in the IV. Under these conditions, drivers were informed by the on-board research staff member when to start and finish the approaching manoeuvres. The order, the starting location, and the extension of the driving

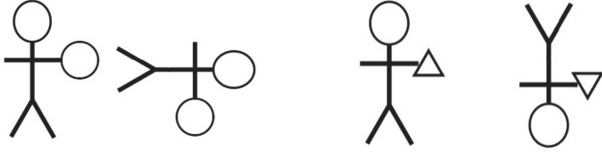


FIGURE 1: Examples of same (left) and different (right) trials for the rotated figures tasks.

conditions were the same for all participants, in both the two experimental environments. The data collection lasted two months and was carried out in daily experimental tests, each consisting of a few (up to 5) driving tests.

All drivers experienced the same experimental conditions, both on the road and at the simulator. It is worth noting that the driving simulator was provided with a traffic-simulation module that allowed traffic conditions to be emulated according to the desired average speed of the traffic stream and traffic density. These parameters were fixed to adapt experimental conditions at the simulator with prevailing ones on the road. Weather conditions did not differ substantially across participants, nor did traffic conditions. Starting times of driving tests were scheduled every two hours, to accommodate the driving time and the time required to complete the questionnaires described in the following section. The order of the simulator and on-road drive conditions was randomized for each participant.

3.3. Mental Workload and Psychological Measures. Rotated figures task (RFT, as in [21]): it is a self-paced secondary task used to measure performance-based workload. It has been successfully adopted in previous research about workload assessment as a secondary task during driving and its reliability can compensate for its apparent lack of ecological validity. To ensure the effective measure of the spare attentional capacity, this subsidiary task is designed to load the same attentional resources as driving (which are visual input, spatial processing, and manual response [14]). Each task stimulus was a pair of stick figures (one upright; the other rotated through 0° , 90° , 180° , or 270°) holding one or two simple geometrical shapes (either triangles, circles, or diamonds; Figure 1).

Participants were asked to perform three dual-task sessions; in each one a series of secondary rotated figures task was assigned in concurrency with the primary driving task. The sessions took place in each of the three driving conditions/scenarios and consisted, respectively, of 9, 13, and 10 trials. During each trial, which was announced by a sound, each pair of figures was directly presented for 15 seconds on the simulator screen in the DS and in the dedicated monitor in the IV. To minimize the eye-off-the-road time [42], the stimulus was presented just above the speedometer in the middle console, in both the IV (in the touch-screen panel PC) and the DS (in a virtual screen on a portion of the central monitor). In either case, participants provided their answer by pressing one of the two push-buttons (left=same, right=different) positioned symmetrically on the steering

wheel with the thumb of their closer hand. Interstimulus intervals could randomly range from 15 to 25 seconds. To avoid the disruption of the primary task performance [43] and thus jeopardize their safety during the on-road drive, participants were instructed to give maximal attention to the primary task, attending to the secondary task only when they felt that they had time to do so in both the test environments. Moreover, a driving instructor on board was always in the control loop of the driving task and able to recall the driver in the case of decreased attention on the primary task, as well as terminate the secondary task. Each dual-task session was performed on the same road section in the two experimental environments.

We then computed accuracy, i.e., rate of correct discriminations (ACC) and average response time (ART). However, these two measures can provide only partial information about performance, and they can conflate experimental factors with strategic effects employed by the participant. Their known covariation (i.e., faster response time is associated with lower accuracy, and vice versa) has traditionally been seen as a signature of the decision process that leads to the final answer. A measure that quantifies precisely how accuracy trades off with latency is thus a useful add-on [44]. Hence, a combined time/correctness factor (CTCF) was also computed. The CTCF combines, for each participant j , response time and accuracy, by the formula given in

$$\alpha_j = \sum_{i=1}^N \frac{e_i \cdot (15 - t_{ieff})}{N} \quad (1)$$

where e_i is the result of the i -th response (1 for correct responses, 0 for wrong ones); 15 are the seconds available to give a response; t_{ieff} is the i -th effective response time; N is the number of trials in the specific dual-task session. The resulting measure was then normalized over the whole sample by

$$CTFT = \frac{\alpha_j - \alpha_{min}}{\alpha_{max} - \alpha_{min}} \quad (2)$$

where α_j is the participant score, α_{min} the smallest α value observed, and α_{max} the largest α value observed.

NASA-Task Load Index (NASA-TLX; [45]; Italian version in [46]): The NASA-TLX is a subjective measure of mental workload. At the end of the driving tasks, participants first rated on a 20-point rating scale six sources of workload: Mental Demand (amount of mental and perceptual activity required to perform the task), Temporal Demand (the amount of time pressure felt by the driver due to the pace at which the tasks or task elements occurred), Physical Demand (the amount of physical activity required), Effort (how hard the driver had to work to accomplish her/his level of performance), Performance (the level of dissatisfaction with the performance), and Frustration (the extent to which the driver felt irritated, stressed, or annoyed). Next, the participant was asked to choose which source contributed more than the other to workload in all the 15 possible pairwise comparisons of workload sources. An aggregated overall workload score was also computed for the simulator and the road condition.

Short Stress State Questionnaire (SSSQ; [47]): The SSSQ is a 24-item multidimensional self-report measure of stress state derived from the Dundee Stress State Questionnaire [48] that provides a quick assessment of three broad higher-order stress state factors, Distress (DI, unpleasant mood and tension with lack of confidence and perceived control), Task Engagement (TE, energetic arousal, motivation, and concentration), and Worry (WO, self-focused attention, self-esteem, and cognitive interference). Participants were asked to complete the pretask and posttask versions of the SSSQ in the simulator and road environment rating items on a six-point, Likert-type intensity scale. The SSSQ was translated independently by two of the authors and by two PhD students fluent in English. When they reached a consensus translation, it was checked through back-translation by an English mother-tongue professional translator. Preliminary analyses showed that the SSSQ maintained the original three-factor structure also in the Italian version. In this study reliabilities (Cronbach's Alphas) were $TE_{pre}=0.87$, $TE_{post}=0.81$, $DI_{pre}=0.85$, $DI_{post}=0.89$, $WO_{pre}=0.79$, and $WO_{post}=0.81$ in the simulator and $TE_{pre}=0.86$, $TE_{post}=0.84$, $DI_{pre}=0.81$, $DI_{post}=0.85$, $WO_{pre}=0.80$, and $WO_{post}=0.79$ on the road. Scale reliabilities did not significantly differ across conditions.

4. Results

Relative validity was examined by computing correlations between scores of the same task performance indices or scale scores in the simulator and road environment. Grounding on previous studies (i.e., [31, 49, 50]), we expected that correlation coefficients could provide from weak (lower than .20) to strong (higher than .50) support to relative validity, depending on the measure being considered. Absolute validity was examined by comparing mean scores across driving conditions using analysis of variance (ANOVA) models. Since the significance of such test can depend on sample size (i.e., negligible effects can be statistically significant in large samples, whereas large effects may not be statistically significant in small samples), we also computed the effect size, i.e., a quantitative measure of the strength of the effect. For the sake of simplicity, we used the same metric as the correlation coefficients mentioned above. Data from thirteen participants were not included in the analyses since they reported simulator sickness. Hence, the sample on which the statistical analyses were carried out comprised 87 cases, with no missing data.

4.1. Behavioral Measure of Mental Workload (Rotated Figures Task)

4.1.1. Accuracy. A completely within-subject 3 (sessions) \times 2 (environments: road vs. simulator) factorial analysis of variance (ANOVA) model showed that all effects were significant (session: $F(2,172)=18.05$, $p<0.001$, $r=0.20$ (r is a standardized measure of effect size and can be interpreted as follows: $r<0.10$ negligible effect; $0.10<r<0.30$ small effect, $0.30<r<0.50$ moderate effect, $r>0.50$ large effect [51]); environment: $F(1,86)=71.58$, $p<0.001$, and $r=0.37$; interaction:

$F(2,172)=52.34$, $p<0.001$, $r=0.32$) (Figure 2(a)). Bonferroni-corrected post hoc tests revealed that in dual-task session 3 accuracy was higher than in the other two sessions on the road, whereas it was lower than in the other two sessions in the simulator. In general, accuracy was significantly higher in the road environment. Correlations across environments of session accuracies ranged from 0.41 to 0.50 and they were all statistically significant at $p<0.05$.

4.1.2. Response Times. A completely within-subject 3 (sessions) \times 2 (environment: road vs. simulator) factorial analysis of variance (ANOVA) model showed that both main effects (session: $F(2,172)=25.76$, $p<0.001$, $r=0.19$; environment: $F(1,86)=12.50$, $p<0.001$, $r=0.17$) were significant, whereas the interaction was not ($F(2,172)=1.08$, $p=0.344$, $r=0.04$) (Figure 2(b)). Bonferroni-corrected post hoc tests revealed that RTs decreased linearly from session 1 to session 3 and that they were generally higher on the road. Correlations across environments of session RTs ranged from 0.47 to 0.55 and they were all statistically significant at $p<0.05$.

4.1.3. Combined Time/Correctness Factor. A completely within-subject 3 (sessions) \times 2 (environment: road vs. simulator) factorial analysis of variance (ANOVA) model showed that the main effect of session ($F(2,172)=5.32$, $p=0.006$, $r=0.10$) and the interaction ($F(2,172)=3.42$, $p=0.035$, $r=0.08$) were significant, whereas the main effect of environment was not significant ($F(1,86)=0.03$, $p=0.873$, $r=0.01$) (Figure 2(c)). Bonferroni-corrected post hoc tests revealed that CTCFs were generally higher in dual-task session 3 than in the other two sessions and that this same pattern could be observed on the road but not in the simulator. Correlations across environments of session CTCFs ranged from 0.50 to 0.63 and they were all statistically significant at $p<0.05$.

4.2. Subjective Measure of Mental Workload (NASA-TLX). A completely within-subject 6 (NASA-TLX scale) \times 2 (environment: road vs. simulator) factorial analysis of variance (ANOVA) model showed that both main effects (scale: $F(5,430)=67.89$, $p<0.001$, $r=0.60$; environment: $F(1,86)=18.89$, $p<0.001$, $r=0.25$) were significant, whereas the interaction was not ($F(2,172)=1.05$, $p=0.389$, $r=0.09$) (Figure 3).

Bonferroni-corrected post hoc tests revealed that in general Mental Demand and Performance were the workload facets that received the highest weight, whereas Frustration was the weakest contributor to workload. In general, simulator scores tended to be higher, but when corrected for multiple comparisons, differences in Temporal Demand and Performance were no longer statistically significant. Correlations across environments of NASA-TLX scale scores ranged from 0.22 to 0.51 and they were all statistically significant at $p<0.05$.

Differences in NASA-TLX total scores were tested through a paired-t test, which showed that scores in the simulator were statistically higher than in the on-road condition (39.51 ± 18.45 vs. 30.86 ± 15.22 , $t(86)=4.35$, $p<0.001$, $r=0.43$). The correlation among the two NASA-TLX total scores was 0.41 ($p<0.001$).

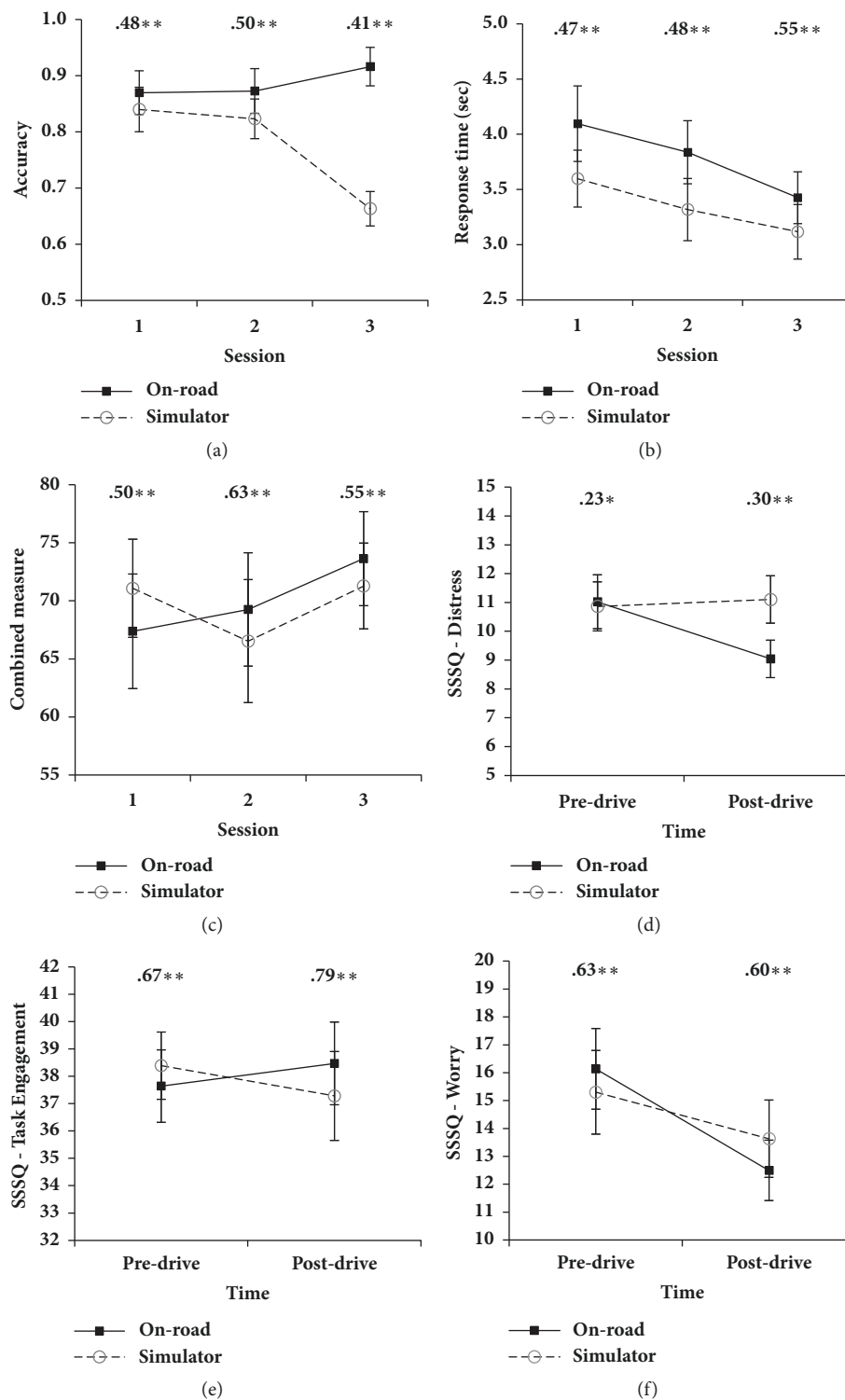


FIGURE 2: On-road vs. simulator mean score comparisons of behavioral measures of the workload (rotated figures task scores: (a) accuracy; (b) response time; (c) combined time/correctness measure) across the three dual-task sessions and of stress measures before and after drive (Short Stress State Questionnaire [SSSQ] scores: (d) Distress, (e) Task Engagement, (f) Worry). Figures inside the graph are Pearson's zero-order correlations between the scores in the simulator and scores on the road as an index of relative validity ($n = 87$); *: $p < 0.05$ and **: $p < 0.01$.

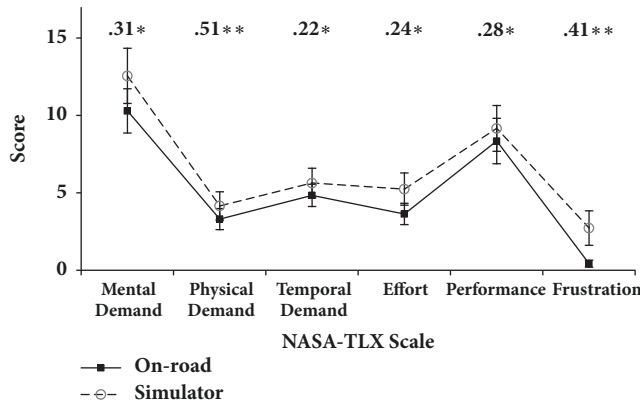


FIGURE 3: On-road vs. simulator mean score comparisons of subjective measures of workload (NASA-Task Load Index scores). Figures inside the graph are Pearson's zero-order correlations between the scores in the simulator and scores on the road as an index of relative validity ($n = 87$). *: $p < 0.05$ and **: $p < 0.01$.

4.3. Short Stress State Questionnaire. A completely within-subject 2 (time: before vs. after task) \times 2 (environment: road vs. simulator) factorial analysis of variance (ANOVA) model showed that for Distress scores all effects were significant (environment: $F(1,86)=8.12$, $p=0.005$, $r=0.16$; time: $F(1,86)=5.29$, $p=0.024$, $r=0.15$; interaction: $F(1,86)=7.67$, $p=0.007$, $r=0.17$) (Figure 2(d)). Bonferroni-corrected post hoc tests revealed that before the driving task Distress scores did not differ across environments, whereas they significantly decreased after driving on the road. Correlations across environments of Distress scores were 0.23 before task and 0.30 after task, and they were all statistically significant at $p < 0.05$.

The same analysis showed that for Task Engagement scores only the interaction effect was significant (environment: $F(1,86)=0.24$, $p=0.626$, $r=0.02$; Time: $F(1,86)=0.06$, $p=0.815$, $r=0.01$; interaction: $F(1,86)=8.22$, $p=0.005$, $r=0.08$) (Figure 2(e)). Bonferroni-corrected post hoc tests revealed that before the driving Task Engagement scores did not differ across environments, whereas they were higher on the road after the task. Correlations across environments of Task Engagement scores were 0.67 before task and 0.79 after task, and they were all statistically significant at $p < 0.05$.

The same analysis showed that for Worry scores the main effects of time and the interaction were significant (environment: $F(1,86)=0.20$, $p=0.654$, $r=0.02$; time: $F(1,86)=30.16$, $p < 0.001$, $r=0.24$; interaction: $F(1,86)=6.57$, $p=0.012$, $r=0.09$) (Figure 2(f)). Bonferroni-corrected post hoc tests revealed that before the driving task Worry scores did not differ across environments whereas they were higher in the simulator after the task. Correlations across environments of Worry scores were 0.63 before task and 0.60 after task, and they were all statistically significant at $p < 0.05$.

5. Discussion

This study investigated the validity of a DS-based experimental environment for research on mental workload by

comparing a number of workload measures of the same group of participants in a simulated and in an on-road driving task on the same route. The findings of this study provided mixed support for both the absolute and relative validity of the simulator in terms of workload. Consistent with the literature (e.g., [31, 38, 39, 49, 50]), some measures showed adequate similarities between the simulated and the real scenario, whereas others did not.

In general, simulated driving led to higher workload and stress levels, somewhat counterintuitively, given the higher actual risk in real traffic, but these differences are not clear-cut and need to be considered in detail. Accuracy in a secondary task tended to be higher (lower workload) on the road (Figure 2(a)), but, interestingly, the difference was higher in the third session, corresponding to less demanding section (Section 3) of the route: while the accuracy increased in the road condition, it decreased in the simulator condition, suggesting a higher workload. However, the higher accuracy in the road condition did not come at zero cost, since response times were significantly higher (more workload) than in the simulator condition and linearly decreased as the driving task progressed (Figure 2(b)). When the performance measure that combined both accuracy and response time was considered (Figure 2(c)), a different pattern of results was found between the road and the simulator conditions: while in session 3 scores were higher (less workload) than in the other sessions of the road condition, no significant differences across sessions were found in the simulator condition. It must be noted that another explanation for these results might be that the driving simulation task tended to be more tiring for the participant (as also shown by the subjective ratings), thus accounting for the decrease in accuracy of the secondary task with time. This would be consistent with the results of the combined measure, which was higher in the simulation condition at the beginning but it tended to be higher in the road condition as the experimental task progressed.

When subjective workload was taken into account, Mental Demand (the amount of mental and perceptual activity required) and performance (the level of dissatisfaction with the performance) received the higher scores, whereas Frustration (the extent to which the driver felt irritated, stresses, and annoyed) received the lowest scores (Figure 3). The pattern of results was basically the same across road and simulator environments, although scores in the simulator tended to be higher. Stress scores showed that although the scale scores did not significantly differ across environments before the task, Distress and Worry scores were lower and Task Engagement scores were higher on the road after driving (Figures 2(d)-2(f)). Although these results seem to suggest a limited absolute validity of the simulator, it can be argued that the higher workload in the simulator can be the result of the lack of familiarity of the participants with the simulator environment and controls and thus the need to learn a "new" behavior. This shortcoming can be addressed by allowing more and/or longer sessions of training [52], also known to reduce simulation sickness [53]. From a relative validity point of view, correlations among similar tasks and measures across road and simulator conditions showed an adequate correspondence of behavioral measures

and of subjective measures of Task Engagement and Worry, as correlations ranged from .40 to .80. Correlations among the other measures, especially Distress and subjective workload ones, seemed to provide only a limited support to relative validity (though they were statistically significant and in the expected direction) but the fact that these measures might be affected by situational and individual factors that are almost uncontrollable and can attenuate the correlations must be taken into account.

Despite the effort to collect a sample of participants that was representative of the Italian driver population, data from thirteen cases could not be used in statistical analyses since participants reported simulator sickness, thus limiting the generalizability of results. Consistent with the literature [54], they were mostly women (62% of all dropouts) and older participants (median age 60 years, range 34-77). Although some solutions have been proposed [55], this common side effect appears to be inevitable in simulation studies, but it must be taken into account when the aim of the simulation study is to test in-vehicle systems and related human-machine interfaces aiming to reduce the workload. In fact, elder drivers are a category that can benefit more than others from these devices, and not being able to generalize research results to this subpopulation would be problematic.

6. Conclusions

The study presented in this paper is part of an on-going research project that aims to gain insights into the level of immersion needed to elicit the desired presence during driving simulation and the interaction between individual differences (e.g., gender, age, personality traits, and driving attitudes) and driving performance measures (e.g., speeding, lane changes, and steering behavior) in both on-road and simulated driving conditions, in order to further test the functional fidelity [12] of driving-simulator-based experimental environments. The present study provided some empirical support for the validity of a fixed-based driving simulator as a safe method of assessing mental workload during driving, although it did not completely clarify whether differences with respect to the on-road condition could be due to insufficient familiarization with the simulator. Simulator sickness also emerged as a critical issue for the generalizability of the results. However, future research is necessary, to understand the underlying mechanisms of these effects.

Data Availability

The DRIVE IN² project data used to support the findings of this study could be made available by contacting Prof. Gennaro N. Bifulco, gennaro.bifulco@unina.it.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors wish to thank Mikaela McKellar and Mark Walters for proofreading their paper. Data were collected within the Italian PON (Programma Operativo Nazionale) 2007-13, Research Project B61H110004000005 *DRIVE IN*².

References

- [1] P. S. Tsang and G. F. Wilson, "Mental workload," in *Handbook of human factors*, G. Salvendy, Ed., pp. 417–449, John Wiley and Sons, New York, NY, USA, 1997.
- [2] C. D. Wickens and J. S. McCarley, *Applied attention theory*, CRC Press, New York, 2008.
- [3] G. Matthews, D. J. Saxby, G. J. Funke, A. K. Emo, and P. A. Desmond, "Driving in states of fatigue or stress," in *Handbook of Driving Simulation for Engineering, Medicine, and Psychology*, and Psychology, D. L. Fisher, M. Rizzo, J. Caird, and J. D. Lee, Eds., pp. 29–11, Taylor and Francis, Boca Raton, FL, USA, 2011.
- [4] D. de Waard and K. Brookhuis, "On the measurement of driver workload," in *Traffic and Transport Psychology: Theory and Application*, T. Rothengatter and E. E. Carbonell Vaya, Eds., pp. 161–171, Amsterdam, Pergamon, 1997.
- [5] Connectivity and Automated Driving, "Automated Driving Roadmap," ERTRAC Working Group, 2017.
- [6] C. Neubauer, G. Matthews, L. Langheim, and D. Saxby, "Fatigue and voluntary utilization of automation in simulated driving," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 54, no. 5, pp. 734–746, 2012.
- [7] K. E. Adanu and S. Jones, "Effects of Human-Centered Factors on Crash Injury Severities," *Journal of Advanced Transportation*, Article ID 1208170, 2017.
- [8] Adrian Fazekas, Friederike Hennecke, Eszter Kalló, and Markus Oeser, "A Novel Surrogate Safety Indicator Based on Constant Initial Acceleration and Reaction Time Assumption," *Journal of Advanced Transportation*, vol. 2017, Article ID 8376572, 9 pages, 2017.
- [9] C. M. Rudin-Brown, A. Williamson, M. G. Lenné, and M. G. Lenné, "Can driving simulation be used to predict changes in real-world crash risk?" Report n. 299, Monash University Accident Research Centre, Victoria, Australia, 2009.
- [10] W. F. Moroney and M. G. Lilienthal, "Human factors in simulation and training: An overview," in *Human Factors in Simulation and Training*, P. Hancock, D. Vincenzi, J. Wise and, and M. Mouloua, Eds., p. 38, Francis Group, LLC, Boca Raton, Taylor, 2009.
- [11] S. Espié, P. Gauriat, M. Duraz, and S. Espié, "Driving simulators validation: the issue of transferability of results acquired on simulator," in *Proceedings of the Driving Simulation Conference - North America*, Orlando, FL, USA, 2005.
- [12] G. Matthews, J. S. Warm, L. E. Reinerman-Jones et al., "The functional fidelity of individual differences research: The case for context-matching," *Theoretical Issues in Ergonomics Science*, vol. 12, no. 5, pp. 435–450, 2011.
- [13] S. Zhao, W. Guo, and C. Zhang, "Extraction Method of Drivers Mental Component Based on Empirical Mode Decomposition and Approximate Entropy Statistic Characteristic in Vehicle Running State," *Journal of Advanced Transportation*, Article ID 9509213, 2017.
- [14] C. D. Wickens, "Multiple resources and performance prediction," *Theoretical Issues in Ergonomics Science*, vol. 3, no. 2, pp. 159–177, 2002.

- [15] C. D. Wickens and J. G. Hollands, *Engineering Psychology*, Prentice-Hall, Upper Saddle River, NJ, 1999.
- [16] G. F. Wilson and F. T. Eggemeier, "Psychophysiological assessment of workload in multi-task environments," in *In Multiple-Task Performance*, Performance. Multiple-Task and D. L. Damos, Eds., pp. 279–328, Taylor Francis, London, 1991.
- [17] K. A. Brookhuis and D. de Waard, "Monitoring drivers' mental workload in driving simulators using physiological measures," *Accident Analysis & Prevention*, vol. 42, no. 3, pp. 898–903, 2010.
- [18] H. Wiberg, E. Nilsson, P. Lindén, B. Svanberg, and L. Poom, "Physiological responses related to moderate mental load during car driving in field conditions," *Biological Psychology*, vol. 108, pp. 115–125, 2015.
- [19] A. Stuiver, K. A. Brookhuis, D. de Waard, and B. Mulder, "Short-term cardiovascular measures for driver support: Increasing sensitivity for detecting changes in mental workload," *International Journal of Psychophysiology*, vol. 92, no. 1, pp. 35–41, 2014.
- [20] M. H. Martens and W. Van Winsum, "Measuring Distraction: the Peripheral Distraction Task, NHTSA Internet Forum on the Safety Impact of Driver Distraction when Using In-Vehicle Technologies," *Measuring Distraction: the Peripheral Distraction Task, NHTSA Internet Forum on the Safety Impact of Driver Distraction when Using In-Vehicle Technologies*, 2000.
- [21] N. A. Stanton, M. Young, and B. McCaulder, "Drive-by-wire: The case of driver workload and reclaiming control with adaptive cruise control," *Safety Science*, vol. 27, no. 2–3, pp. 149–159, 1997.
- [22] P. S. Tsang and V. L. Velazquez, "Diagnosticity and multidimensional subjective workload ratings," *Ergonomics*, vol. 39, no. 3, pp. 358–381, 1996.
- [23] Yei-Yu Yeh and C. D. Wickens, "Dissociation of performance and subjective measures of workload," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 30, no. 1, pp. 111–120, 1988.
- [24] M. A. Vidulich and C. D. Wickens, "Causes of dissociation between subjective workload measures and performance. Caveats for the use of subjective assessments," *Applied Ergonomics*, vol. 17, no. 4, pp. 291–296, 1986.
- [25] W. J. Horrey, M. F. Lesch, and A. Garabet, "Dissociation between driving performance and drivers' subjective estimates of performance and workload in dual-task conditions," *Journal of Safety Research*, vol. 40, no. 1, pp. 7–12, 2009.
- [26] A. Brandes, M. D. Smit, B. O. Nguyen, M. Rienstra, and I. C. Van Gelder, "Risk Factor Management in Atrial Fibrillation," *Arrhythmia & Electrophysiology Review*, vol. 7, no. 2, p. 118, 2018.
- [27] C. J. D. Patten, A. Kircher, J. Östlund, and L. Nilsson, "Using mobile telephones: Cognitive workload and attention resource allocation," *Accident Analysis & Prevention*, vol. 36, no. 3, pp. 341–350, 2004.
- [28] G. Jahn, A. Oehme, J. F. Krems, and C. Gelau, "Peripheral detection as a workload measure in driving: effects of traffic complexity and route guidance system use in a driving study," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 8, no. 3, pp. 255–275, 2005.
- [29] F. A. Muckler and S. A. Seven, "Selecting performance measures: 'Objective' versus 'subjective' measurement," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 34, no. 4, pp. 441–455, 1992.
- [30] G. J. Blaauw, "Driving experience and task demands in simulator and instrumented car - a validation study," *Hum Factors*, vol. 24, no. 4, pp. 473–486, 1982.
- [31] S. T. Godley, T. J. Triggs, and B. N. Fildes, "Driving simulator validation for speed research," *Accident Analysis & Prevention*, vol. 34, no. 5, pp. 589–600, 2002.
- [32] M. P. Reed and P. A. Green, "Comparison of driving performance on-road and in a low-cost simulator using a concurrent telephone dialling task," *Ergonomics*, vol. 42, no. 8, pp. 1015–1037, 1999.
- [33] J. Törnros, "Driving behaviour in a real and a simulated road tunnel - A validation study," *Accident Analysis & Prevention*, vol. 30, no. 4, pp. 497–503, 1998.
- [34] N. A. Kaptein, J. Theeuwes, and R. van der Horst, "Driving simulator validity: some considerations," *Transportation Research Record*, no. 1550, pp. 30–36, 1996.
- [35] F. Galante, F. Mauriello, A. Montella, M. Perneti, M. Aria, and A. D'Ambrosio, "Traffic calming along rural highways crossing small urban communities: Driving simulator experiment," *Accident Analysis & Prevention*, vol. 42, no. 6, pp. 1585–1594, 2010.
- [36] N. Stanton, M. S. A. G. H. Walker, A. Turner, and S. Randle, "Automating the driver's control tasks," in *Proceedings of the Automating the driver's control tasks. International. Journal of Cognitive Ergonomics*, vol. 5, pp. 221–236, 2001.
- [37] I. D. Brown, "Dual task methods of assessing work-load," *Ergonomics*, vol. 21, no. 3, pp. 221–224, 1978.
- [38] M. J. Johnson, T. Chahal, A. Stinchcombe, N. Mullen, B. Weaver, and M. Bédard, "Physiological responses to simulated and on-road driving," *International Journal of Psychophysiology*, vol. 81, no. 3, pp. 203–208, 2011.
- [39] Y. Wang, B. Mehler, B. Reimer, V. Lammers, L. A. D'Ambrosio, and J. F. Coughlin, "The validity of driving simulation for assessing differences between in-vehicle informational interfaces: a comparison with field testing," *Ergonomics*, vol. 53, no. 3, pp. 404–420, 2010.
- [40] L. Pariota, G. N. Bifulco, F. Galante, A. Montella, and M. Brackstone, "Longitudinal control behaviour: Analysis and modelling based on experimental surveys in Italy and the UK," *Accident Analysis & Prevention*, vol. 89, pp. 74–87, 2016.
- [41] G. N. Bifulco, L. Pariota, F. Galante, and A. Fiorentino, "Coupling instrumented vehicles and driving simulators: Opportunities from the DRIVE IN2 project," in *Proceedings of the 2012 15th International IEEE Conference on Intelligent Transportation Systems, ITSC 2012*, pp. 1815–1820, USA, September 2012.
- [42] M. Wittmann, M. Kiss, P. Gugg et al., "Effects of display position of a visual in-vehicle task on simulated driving," *Applied Ergonomics*, vol. 37, no. 2, pp. 187–199, 2006.
- [43] H. A. Colle and G. B. Reid, "Double trade-off curves with different cognitive processing combinations: Testing the cancellation axiom of mental workload measurement theory," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 41, no. 1, pp. 35–50, 1999.
- [44] R. P. Heitz, "The speed-accuracy tradeoff: History, physiology, methodology, and behavior," *Frontiers in Neuroscience*, no. 8, 2014.
- [45] S. G. Hart and L. E. Staveland, "Development of a multi-dimensional workload rating scale: Results of empirical and theoretical research," in *Human mental workload*, P. A. Hancock and N. Meshkati, Eds., pp. 139–183, Elsevier, Amsterdam, 1988.
- [46] F. Bracco and C. Chiorri, "Validazione italiana del NASA-TLX su un campione di motociclisti [Italian validation of the NASA-TLX on a sample of bikers]," in *Proceedings of the XV Congresso Nazionale dell'Associazione Italiana di Psicologia, Sezione di Psicologia Sperimentale*, Rovereto, Italy, 2006.

- [47] W. S. Helton, "Validation of a Short Stress State Questionnaire," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 48, no. 11, pp. 1238–1242, 2016.
- [48] G. Matthews, L. Joyner, K. Gilliland, J. Huggins, and S. Falconer, "Validation of a comprehensive stress state questionnaire: Towards a state big three?" in *Personality psychology in Europe*, I. Merville, I. J. Deary, F. DeFruyt, and F. Ostendorf, Eds., vol. 7, pp. 335–350, Tilburg University Press, 1999.
- [49] H. Devos, A. Nieuwboer, W. Vandenberghe, M. Tant, W. De Weert, and E. Uc, "Validation of driving simulation to assess on-road performance in Huntington disease," in *Proceedings of the Seventh International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design*, pp. 241–247, Bolton Landing, New York, USA, 2013.
- [50] D. R. Mayhew, H. M. Simpson, K. M. Wood, L. Lonero, K. M. Clinton, and A. G. Johnson, "On-road and simulated driving: Concurrent and discriminant validation," *Journal of Safety Research*, vol. 42, no. 4, pp. 267–275, 2011.
- [51] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, Lawrence Erlbaum Associates, New York, NY, USA, 2nd edition, 1988.
- [52] M. Klüver, C. Herrigel, C. Heinrich, H.-P. Schöner, and H. Hecht, "The behavioral validity of dual-task driving performance in fixed and moving base driving simulators," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 37, pp. 78–96, 2016.
- [53] N. Teasdale, M. Lavallière, M. Tremblay, D. Laurendeau, M. Simoneau, and M. Lavallière, "Multiple exposition to a driving simulator reduces simulator symptoms for elderly drivers," in *Proceedings of the Fifth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, pp. 169–175, Big sky, Montana, USA, 2009.
- [54] S. Classen, M. Bewernitz, and O. Shechtman, "Driving simulator sickness: An evidence-based review of the literature," *American Journal of Occupational Therapy*, vol. 65, no. 2, pp. 179–188, 2011.
- [55] H. B.-L. Duh, D. E. Parker, and T. A. Furness, "An independent visual background reduced simulator sickness in a driving simulator," *Presence: Teleoperators and Virtual Environments*, vol. 13, no. 5, pp. 578–588, 2004.

